CD/SCF/DMS/DMD

Enstore SFA Migration Phase 2 Design Document

Alex Kulyavtsev

Date	6/25/14
Version	0.2

Table of Contents

Introduction

Small File Aggregation feature (SFA) of Enstore tape system provides transparent aggregation of small files into larger containers (packages) written to tapes. Migration and Duplication features of Enstore provide capabilities to migrate data from old generation of tapes to the newer technology tapes or create duplicate (mirror) copy of the files in different tape library for better reliability.

The migration and duplication in the initial implementation of SFA is performed by migration and duplication of physical files on tape 'as is' and proper modification of file metadata in enstore DB. This allows migrating file packages to new tape technology generation. We have large amounts of data written to tapes (LTO4) before SFA was introduced. Some datasets have "small" files yet unpackaged that need to be *packaged*. For the files written with and packaged by SFA the new tape technologies may offer better performance if the physical file (package) written to the tape is larger. To enable file *repackaging* in SFA migration process shall unwind packages and write them to the new tape using new policy.

The phase two of SFA Migration and Duplication will allow to *package* small files and/or to *repackage* packaged SFA files according policy during migration process.

We do not change duplication to repackage files: duplicated files stay packaged the same way both on original and copy tape.

Background information

Migration Operation

In current implementation Migration process reads files from source tape(s) and accumulates files on migration spool area on local disk. There are several (typically three) reader threads to provide parallelism in accessing and checking file metadata while some other process is reading data from tape. The other few (usually three) threads write data to tape from disk spool area, one actually writes ant the other two performing metadata operations or waiting for IO. The threads communicate trough python queues. All migration activity runs on the same node (migration station). Few migration stations operate in parallel working on different set of tapes.

The migration process operates 'physical' file on tape: unpackaged files and packages. Migration process automatically identifies if the file is a package. In this case it creates proper metadata in enstore DB tables for packaged files on new tape. File Clerk performs swap metadata operation at once for all files in the package. This metadata 'swapping' happens during tape 'scan' step. The administrator may decide to reverse result of migration and perform 'restore' operation on migrated file. File metadata are swapped back. The log of ongoing migration operation and history are kept in DB.

SFA operation

Files are written to SFA using encp as usual. If proper conditions are met ("enable-redirection" flag set, file satisfies SFA policy), the file will be written by 'disk mover' to the disk on SFA system, then packaged to tar files and the tar file written to the tape. When reading files packaged files, the container file is read to SFA disk, unpackaged and the small file is delivered through disk mover and encp.

Hardware

Few most recent migration stations have Intel X3360 @2.83GHz 4 core CPU, 8GB memory, 10GbE connection. Migration spool area resides on Nexsan appliance formatted with two 4.5 TB arrays connected through two 8Gbit FC. There is also 367GB SSD currently used as migration spool area.

The current generation of tapes allows to transfer data at speeds 240 MB/sec (T10KC, ~5 TB) or 252 MB/sec (T10KD, 8 TB). We use IO rate=70 MB/sec to estimate "typical" IO rate in enstore to/from LTO-4 800 GB tape.

Proposed Design

Considering our current hardware configuration we suggest to unpack files to the local migration spool disk then write files to enstore through SFA facility.

Commad Line Arguments

Add command line arguments:

`--sfa_repackage`	enable file repackaging. Default: migration by package (no repackaging).

File Reader

Migration of tape

For unpackaged tapes the files on tape are read as usual to the local migration spool area.

Option A) During the tape migration for files written with SFA and packaged the reader will access the full list of files on tape (instead of just physical files) and read files through SFA which will unpack files and deliver them trough encp.

Unpacking full package files on SFA servers can create unneeded load on SFA servers and it can be slow as the chain of communication happens to pull small files through SFA.

Option B) To speedup file reads we will operate physical file on tape only and read full package to the local disk and unwind (e.g. untar) the package to the local migration spool disk. To increase tape IO speed, files will be read to local SSD drive if unpacking spool is configured so. The package files will be read in subdirectory in Migration Spool, and softlink can be set to the directory on SSD.

The option (B) is preferable.

The optimization will be to create named pipe and read SFA package to the named pipe. The unpacker (untar) will read files from the named pipe and unpack selected files using tar argument `--files-from file.list` or all files if all files in the package need to be migrated.

Migration of Individual Packaged Files

It is assumed that migration of the single file by specifying BFID or file name on command line or the file with the list of files is used to “fix” individual files and is not used for large dataset migrations. The reading (and unpacking) of individual packaged files to the local disk is performed trough SFA facility to migration spool area.

Scheduler

The present implementation uses few read threads and few write threads to overlap metadata checks and data IO. It does not check for available disk space on migration spool. We observed situation when files were not removed from spool after the crash and the reader started to fail. The individual files were failed with error and reader rapidly goes through full list.

The reader will be modified to detect `no space left on device` IO errors, and in the first place to check if disk space is available and wait if needed before scheduling read/unpack transfer. It shall account when files are deleted in the spool.

File Writer

Add flag to encp ‘—enable-redirection’ when needed to write data through SFA. We may keep flag –enable-redirection for all files in repackaging run if LMD is modified to check if file is a package and always redirect packages to the tape library.

Implement ‘Flush’ request in SFA to stimulate pushing files to the tape (instead of waiting for timeout expiration).

SFA

Change name for migrated package in SFA.

SFA uses file family name as component of path for the package in /pnfs . Migrated files are written in with file family *file_family*-MIGRAITON thus the word “MIGRAITON” is present in the file path. SFA shall trim word MIGRATION at the end of file family name when creating the package.

Metadata Swapping

Scan

Swap method is called during the scan to swap metadata so pnfs pnfs id and enstore file clerk record pnfs for bfid of the migrated file are properly crossreferenced. Currently we use call to File Clerk to swap all files in the package to the new package at once. This will not work anymore for as the original individual files may be not packaged, or will be repackaged to different packages, thus migration will swap metadata for each file individually.

We do not swap package metadata anymore for packages when repackaging files.

We may swap file metadata as soon as file written to SFA. We assume that SFA will take care of delivering file to tape. It is preferable to run scan as we do now as a separate step. The validation script (not part of the migration) can be amended to check the file was actually written to tape before declaring migration finished.

File Restore

The Restore operation reverses the result of metadata swapping thus it operates in opposite way. Migration will swap metadata for constituent file individually instead of block operation in File Clerk for all files in the package.

Notable Features and Side Effects

When writing files with SFA some files may be packaged and some files are not. Accumulating files in SFA cache takes some time and thus the order of the files on tape will change substantially for some files.

Changes in Tape Migration Procedures

Setup Proper File Aggregation Policy

Add to migration instructions: the proper SFA Policy Engine (PE) policy shall be set by administrators before starting migration to define extra policy for the file family *file_family* -MIGRAITION when migrating files in enstore file family *file_family*.

Migration Validation

SSA performs migration validation using custom scripts. The modification of these scripts is out of scope of this document.

Acronyms

SFA	Small File Aggregation. Enstore feature to write/read small files to tape with transparent packaging/unpackaging.
PE	SFA Policy Engine

References

[1] Enstore Technical Design Document,
http://www-ccf.fnal.gov/enstore/design.html
[2] Fermilab DocDB, CS Document 5035-v2, Enstore Administrator's Guide
[3] Enstore Small File Aggregation HLD,
https://cdcvs.fnal.gov/redmine/documents/58
[4] Fermilab DocDB, CS--doc--4698, Enstore Small File Aggregation Feature User Documentation

Revision History

Document ID	Version	Date	Author	Comments
	1.0	6/25/14	Alex Kulyavtsev	Initial revision